Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 6

1. Pruebas estadísticas

Para realizar algunas las pruebas estadísticas se utilizó la información del Índice Nacional de Precios al Consumidor (INPC)[1] mensual por ciudades en el periodo desde enero 2015 a agosto 2020. Esta información fue consultada en la página del INEGI [3], en la sección de Precios. Se utilizó el programa R versión 4.0.2. [5] basado en la información encontrada en *Statistical Tests* [4] para la ejecución de algunas de las pruebas.

La información de las 55 ciudades del INPC y su respectivo índice con base en la segunda quincena de julio 2018, fueron descargadas en un archivo `xlsx`. Posteriormente, estos datos se guardaron en un archivo `txt` para su tratamiento en el programa R.

Para efectos del estudio, se determinaron dos conjuntos de muestras, el primer conjunto (Conjunto 1) consiste en contemplar solo la información de las 3 principales ciudades de México (Monterrey, Ciudad de México y Guadalajara) por separado. En el segundo (Conjunto 2), se consideró la información de los índices de todas las ciudades en los meses de agosto 2018, julio 2019 y agosto 2020. En todas las pruebas para aceptar la H_0, el valor p debe ser mayor a 0.05 (nivel de significancia α).

1.1. Prueba Shapiro–Wilks

Esta prueba plantea la hipótesis nula (H_0) de que los datos provienen de una distribución normal y una hipótesis alternativa (H_1) que sostiene que la distribución no es normal.

[1] El INPC es un indicador económico que muestra la variación de los precios en un periodo de tiempo.

Cuadro 1: Resultados de la prueba de Shapiro–Wilks aplicada a los Conjuntos 1 y 2

Conjunto	**Contenido**	**Valor** p
1	Monterrey	0.0002
	Cd. México	0.0002
	Guadalajara	0.0007
2	Agosto 2018	0.4888
	Julio 2019	0.0950
	Agosto 2020	0.5525

Se realizó esta prueba con la función `shapiro.test` para determinar, si los datos del Conjunto 1 y 2 siguen una distribución normal. En el cuadro 1, se muestran los valores de p obtenidos para los datos de estos conjuntos. En el cual podemos observar que en el Conjunto 1, el valor p es menor a 0,05 para cada ciudad analizada, por lo tanto, se rechaza la H_0 lo que significa que los datos analizados al parecer no siguen una distribución normal.

No obstante, para el Conjunto 2, los datos de cada mes analizado suponen seguir una distribución normal (valor $p > 0.05$), lo cual podemos observarlo también en la figura 1.

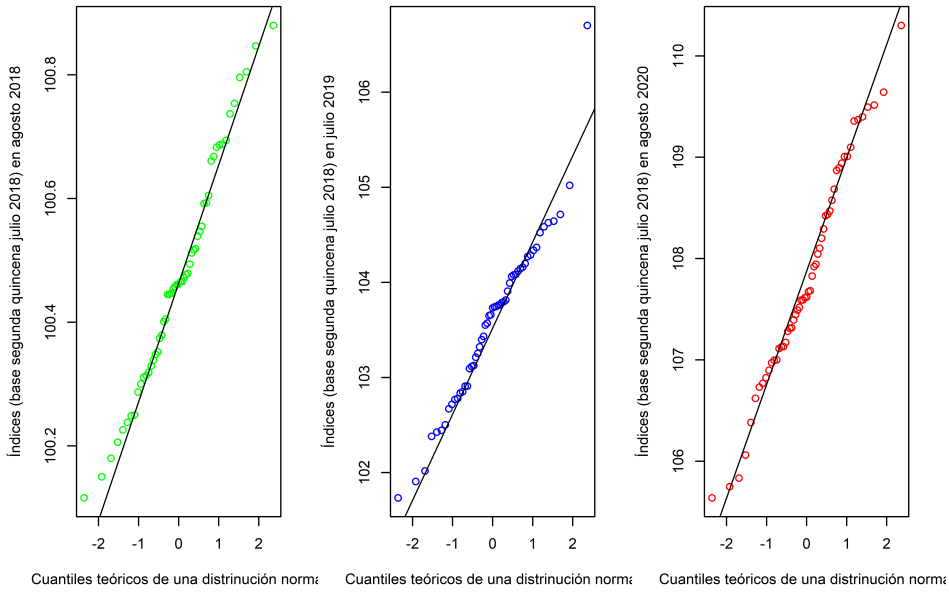

Figura 1: Comparativo de las gráficas QQ Normal de los datos con los índices de las 55 ciudades del INPC en los meses de agosto 2018 (verde), julio 2019 (azul) y agosto 2020 (rojo)

De acuerdo a lo anterior, para efectos de las pruebas que requieren datos con distribución normal, utilizaremos los datos del Conjunto 2 y para las pruebas que requieren que los datos no sigan una

distribución normal, utilizaremos los datos del Conjunto 1.

1.2. Prueba t de student

Esta prueba paramétrica se utiliza para probar si la media (μ) de una muestra con distribución normal es igual a un valor especifico.

Para esta prueba se plantea como H_0 que los datos del mes de agosto 2018 tienen una $\mu = 100.4$ y, como H_1 que estos datos tienen una μ diferente. Se utilizó la función `t.test` para realizar esta prueba. De acuerdo al resultado arrojado por la función, el valor $p = 0.005324$, por lo tanto, se rechaza la H_0, lo cual indica que los datos del mes de agosto 2018 no tienen una media $\mu = 100.4$.

1.3. Prueba de los rangos con signo de Wilcoxon

Esta prueba paramétrica se utiliza para probar si la media (μ) de una muestra, que se supone no sigue una distribución normal, es igual a un valor especifico. Para esta prueba se plantea como H_0 que el promedio de los datos de la ciudad de Monterrey tienen una media de $\mu = 98.5$ y, como H_1 que en promedio estos datos no tienen esta media. Se utilizó la función `wilcox.test` para realizar esta prueba. De acuerdo al resultado arrojado por la función, el valor $p = 0.1779$, por lo tanto, no se rechaza la H_0, lo cual indica que el promedio de los datos de la ciudad de Monterrey supone tener una $\mu = 98.5$.

1.4. Prueba t de student y Wilcoxon para dos muestras

Ambas pruebas se pueden utilizar para comparar la media de 2 muestras. La diferencia es que la prueba t supone que las muestras que se prueban siguen una distribución normal, mientras que para la prueba Wilcoxon se supone que no tienen este tipo de distribución los datos analizados.

Para la prueba de Wilcoxon se plantea como H_0 que los datos de la ciudad de Monterrey y Guadalajara tienen la misma media ($\mu_1 - \mu_2 = 0$) y, como H_1 que sus medias son diferentes ($\mu_1 - \mu_2 \neq 0$). De acuerdo al resultado arrojado por la función, el valor $p = 0.5642$, por lo tanto, no se rechaza la H_0, lo cual indica que las medias de estas dos muestras, al parecer (una probabilidad del 56.42 %), son iguales. Esto significa que en el periodo de enero 2015 a agosto 2020, el INPC de Monterrey y Guadalajara fue similar. Este comportamiento lo podemos observar en la figura 2.

Para la prueba de t se plantea como H_0 que los datos de los meses de agosto de 2018 y agosto 2020 tienen la misma media y, como H_1 que sus medias son diferentes. De acuerdo al resultado arrojado por la función, el valor $p = 2.2\times10^{-16}$, por lo tanto, se rechaza la H_0, lo cual indica que, con una probabilidad muy baja, las medias de estas dos muestras son iguales. En términos del ejercicio, de acuerdo a la figura

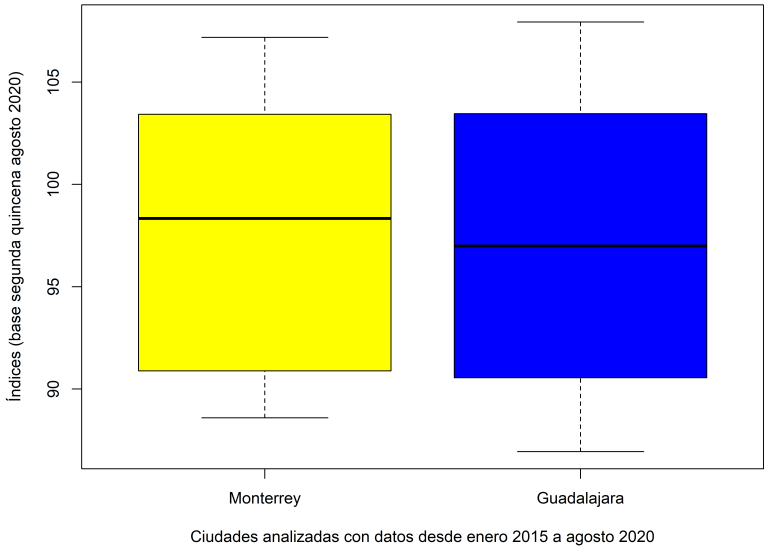

Figura 2: Comparativo de los diagramas de caja y bigotes con los datos recopilados desde enero 2015 a agosto 2018 de los INPC de las ciudades de Monterrey (amarillo) y Guadalajara (azul)

3, el INPC promedio de las 55 ciudades en el mes de agosto de 2018 ($\mu_1 = 100.4723$) fue más bajo que el promedio reportado en agosto de 2020 ($\mu_2 = 107.8254$), es decir, son más costosos los precios de la canasta de bienes y servicios en el mes agosto 2020 que hace 2 años.

1.5. Prueba de Kolmogorov–Smirnov

Esta prueba se utiliza para comprobar si 2 muestras siguen la misma distribución. Para su desarrollo se plantea como H_0 que los datos de Cd. México y Guadalajara tienen una misma distribución y, como H_1 que estas muestras no tienen la misma distribución. Se utilizó la función `ks.test` para realizar esta prueba. De acuerdo al resultado arrojado por la función, el valor $p = 0.7384$, por lo tanto, no se rechaza la H_0, lo cual indica que, al parecer, los datos de Cd. México y Guadalajara tienen una misma distribución.

test.txt

```
        Two-sample Kolmogorov-Smirnov test

data:  cdMexico and guadalajara
D = 0.11765, p-value = 0.7384
alternative hypothesis: two-sided
```

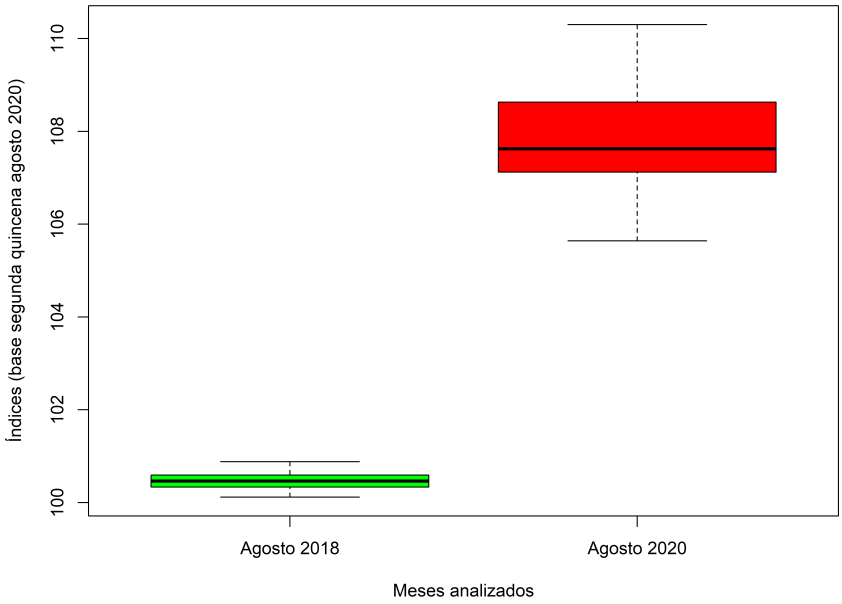

Figura 3: Comparativo de los diagramas de caja y bigotes con los datos recopilados de las 55 ciudades del INPC en los meses agosto 2018 (verde) y agosto 2020 (rojo)

1.6. Prueba F de Fisher

Es una prueba paramétrica que se utiliza para verificar si dos muestras tienen la misma varianza. Se plantea como como H_0 que los datos de los meses de agosto 2018 con julio 2019 tienen la misma varianza y, como H_1 que estas muestras sus varianzas son diferentes. Se utilizó la función `var.test` para realizar esta prueba. De acuerdo al resultado obtenido por la función, el valor $p = 2.2\times10^{-16}$, por lo tanto, se rechaza la H_0, lo cual indica que los datos de los meses de agosto 2018 y julio 2019 no tienen la misma varianza.

test.txt

```
        F test to compare two variances

data:  agosto2018 and julio2019
F = 0.044063, num df = 54, denom df = 54, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.02570083 0.07554563
sample estimates:
ratio of variances
        0.04406342
```

Cuadro 2: Tabla de contingencia para la prueba Chi^2

	Monterrey	Cd. México
Ene 2016	90.779	88.461
Feb 2016	91.292	88.878
Mar 2016	91.400	89.077
Abr 2016	90.261	88.921
May 2016	90.250	88.946
Jun 2016	90.314	88.961
Jul 2016	90.505	89.381
Ago 2016	90.992	89.636
Sep 2016	91.635	90.216
Oct 2016	92.835	90.383
Nov 2016	93.108	90.620
Dic 2016	93.460	91.052

Esta prueba también se puede usar para saber si la varianza de una muestra sigue un determinado valor, solo hay que adicionar ciertos parámetros en la función `var.test` para realizarla. Se pueden ver ejemplos en el repositorio de Freddy [2].

1.7. Prueba Chi^2

Esta prueba se puede utilizar para probar si dos variables categóricas son dependientes, mediante una tabla de contingencia. Para realizar esta tabla, se consideraron dos factores, ciudades y meses del año. Para las ciudades, se contemplan Monterrey y Guadalajara y se analizaran los índices de los 12 meses del año 2016. En el cuadro 2, se encuentra la información de los datoa a analizar.

Se plantea como H_0 que valores del índice alcanzado de cada ciudad es independiente del mes del año en que se encuentren y, como H_0 las variables son dependientes. Se utilizó la función `chisq.test` para realizar esta prueba. Para aceptar la H_0 se deben de cumplir dos consideraciones, que el valor $p > 0.05$ y que el $X - Square$ sea menor al valor crítico. Para calcular este valor se utiliza la función `qchisq(0.95, n-1)`, donde 0.95 es el nivel de confianza y n-1 son los grados de libertad. Estos grados de libertad dependen de la cantidad de filas y columnas que tenga la tabla de contingencia y se calculan con la siguiente formula: $(filas - 1)*(columnas - 1)$. Para este caso, el valor crítico = 19.67514.

Al aplicar la función se obtiene que el valor $p = 1$ y un $X - Square <$ valor crítico, por lo tanto, no hay suficiente evidencia estadística para rechazar H_0, lo cual indica que las variables son independientes, es decir, el índice alcanzado de las ciudades de Monterrey y Guadalajara no dependen del mes del año.

test.txt

```
        Pearson's Chi-squared test

data:  tablacontingencia
X-squared = 0.035179, df = 22, p-value = 1
```

1.8. Prueba de correlación

Es una prueba paramétrica que se utiliza para probar si hay una relación lineal de dos variables continuas. Se pretende determinar si hay una correlación entre los índices de los de meses de julio 2019 y agosto 2020. Se plantea entonces como H_0 que los datos del mes de julio 2019 no están relacionados con el mes de de agosto 2020 y, como H_1 existe una relación entre estas variables. Se emplea la función `cor.test`. De acuerdo a los resultados arrojados por la prueba, el valor $p = 1.136 \times 10^{-9}$, por lo tanto, se rechaza la H_0, lo cual indica que, los datos de los meses de julio 2019 y agosto 2020 tienen alguna correlación. En la figura 4 se muestra el diagrama de dispersión de los datos en el que visualmente se puede observar esta correlación.

test.txt

```
        Pearson's product-moment correlation

data:  julio2019 and agosto2020
t = 7.3722, df = 53, p-value = 1.136e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5500801 0.8217206
sample estimates:
     cor
0.711539
```

2. Preguntas

A continuación, se presentan algunas de las preguntas más frecuentes para determinar que prueba estadística escoger y la interpretación de sus resultados.

- Relación entre contraste de hipótesis y pruebas estadísticas

  Las pruebas estadísticas son técnicas que utilizan muestras representativas de una población para evaluar la evidencia que proporcionan los datos y la hipótesis es la suposición de algún fenómeno

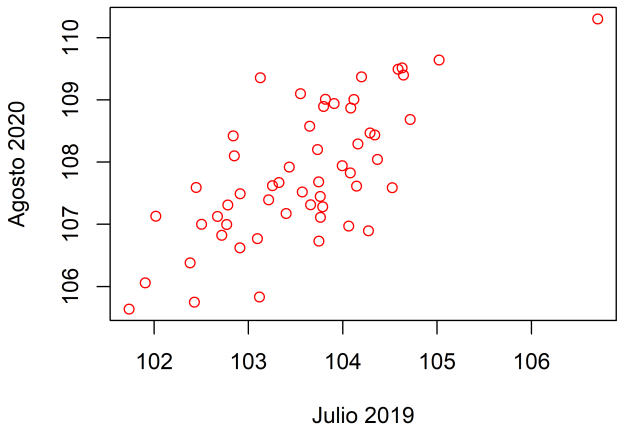

Figura 4: Diagramas de dispersión de los índices en los meses de julio 2019 y agosto 2020

o problema la cual son comprobadas a través de las pruebas estadísticas. En estas pruebas existen dos tipos de hipótesis, nula (H_0) y la alternativa (H_1 o H_A).

- ¿Qué indicaría rechazar la hipótesis nula?

  Normalmente, la hipótesis nula (H_0) establece la igualdad entre las medias, varianzas, entre otros, por lo tanto, si se rechaza indicaría que existe alguna diferencia entre ellas.

- ¿Cómo se interpreta la salida de una prueba estadística?

  Al diseñar un estudio, se especifica un nivel de significación alfa (α) que debería estar entre 0 y 1, lo que indica que por encima este valor H_0 no debería ser rechazada. La prueba estadística produce un número denominado valor p el cual también se encuentra entre 0 y 1. En términos más prácticos, el valor p se compara con α, si $p < \alpha$, rechazamos H_0 y aceptamos Ha con un riesgo proporcional al valor p de ser errónea. Por otro lado, si $p > \alpha$, no rechazamos H_0, pero esto no implica necesariamente que debamos aceptarla, solo que nuestro experimento y nuestra prueba estadística no han sido suficientemente "fuertes"para producir un valor p inferior a α.

- ¿Cómo seleccionar el alpha?

  El nivel de significancia se denota con la letra griega alfa α. No existe una evidencia científica de cuál sea el valor más adecuado. Sin embargo, el valor más empleado es 0.05, se suelen tomar valores pequeños (menores al 10 %), esto debido a que representa el riesgo de rechazar la hipótesis nula (H_0) cuando es verdadera. En ese sentido con una buena elección de α se delimita muy bien cuando rechazar la H_0 lo que aumenta la probabilidad de tomar la decisión correcta.

- ¿Cuáles son los errores frecuentes de interpretación del valor p?

Cuadro 3: Situaciones posibles al probar una hipótesis estadística

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

La interpretación del valor p conduce al rechazo o la aceptación de la hipótesis nula, específicamente el valor p corresponde al menor valor de alfa (α) que ocasiona el rechazo de la hipótesis nula (H_0). Los errores más frecuentes que se presentan son los errores tipo I y tipo II. El error tipo I se presenta cuando se rechaza la H_0 y es verdadera, mientras que el error tipo II consiste en no rechazar la H_0 cuando es falsa. De acuerdo a Walpole [6], en el cuadro 3 se resumen estos errores. La probabilidad de cometer esos errores se reduce aumentando el tamaño de la muestra.

- ¿Qué es la potencia estadística y para qué sirve?

  La potencia estadística o el poder del estadístico, corresponde a la capacidad que tiene una prueba de llevar al rechazo de la hipótesis nula (H_0). La potencia estadística aumenta con el valor de alfa α, con la precisión de las medidas y el número de repeticiones, también depende del tipo de prueba estadística que se esté realizando. Este parámetro puede ser calculado antes o después de realizar el experimento.

- Ejemplos de pruebas estadísticas paramétricas y no paramétricas.

  Las pruebas paramétricas se emplean para datos numéricos, suelen estar basadas en las propiedades de la distribución normal para la variable dependiente. Es decir, los datos son mediciones repetidas de la misma variable, muestreado de la población realizado al azar y cuando la muestra es grande. Como ejemplo de pruebas paramétricas se tienen:

  - La "t" de student.
  - El coeficiente de correlación de Pearson.
  - La regresión lineal.
  - Análisis de varianza unidireccional (ANOVA *Oneway*).
  - Análisis de varianza factorial (ANOVA).
  - Análisis de covarianza (ANCOVA).
  - Estadígrafos descriptivos como la desviación estándar, la moda, la mediana y la media.

  Por otra parte, las pruebas no paramétricas se aplican con variables nominales y ordinales, no asume un tipo específico de distribución. Ejemplos de este tipo de pruebas son:

  - La X^2

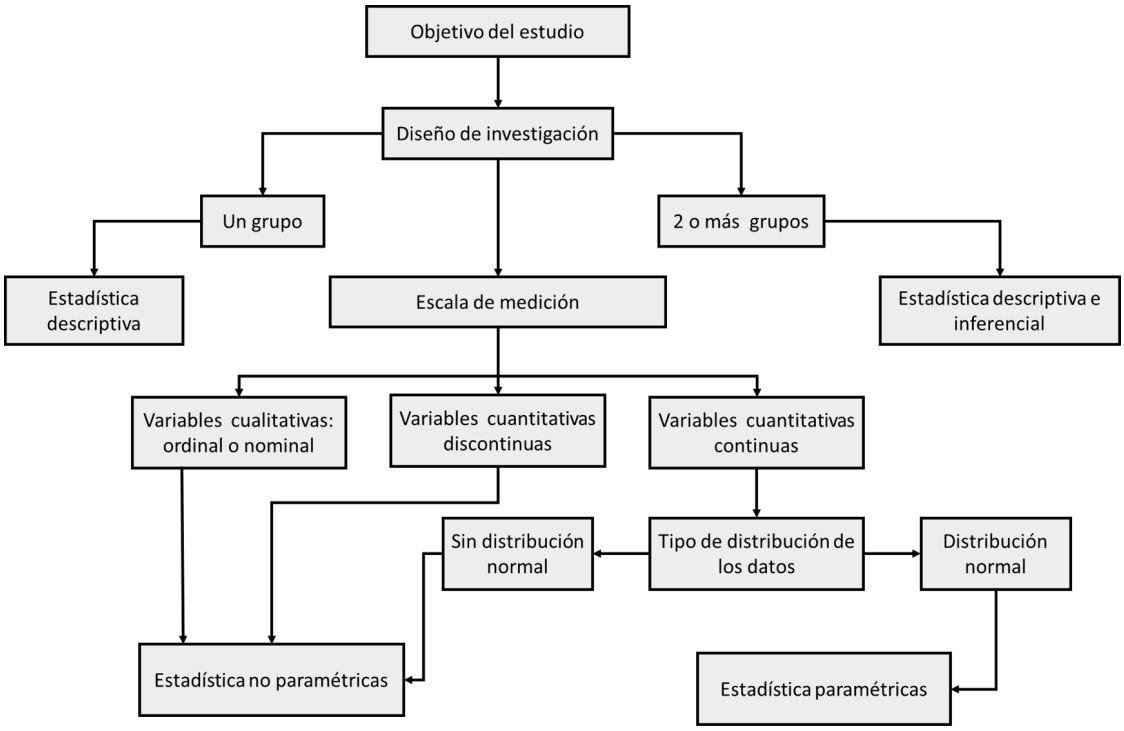

Figura 5: Proceso de selección de una prueba estadística

- Coeficientes de correlación e independencia para tabulaciones cruzadas.
- Coeficientes de correlación por rangos ordenados Spearman y Kendll.

- Resume LA GUIA para encontrar la prueba estadística que buscas.

  - Escribir claramente el objetivo de análisis.
  - Tipo de variables.
  - Si son muestras independientes o no.
  - Identificar si se pueden aplicar técnicas paramétricas.
  - Seleccionar la prueba adecuada.
  - Realizar la prueba de hipótesis.
  - Interpretar y graficar los resultados.

  En la figura 5, se describe el proceso de selección de una prueba estadística basado en el estudio realizado por Florez-Ruiz [1].

- ¿Cuáles son los supuestos para aplicar técnicas paramétricas?

  Las pruebas paramétricas están basadas en la distribución normal para la variable dependiente, y los requisitos para aplicarlas son las siguientes:

- Las observaciones deben ser independientes entre sí.
- Las poblaciones deben hacerse en poblaciones distribuidas normalmente.
- Estas poblaciones deben tener la misma varianza.
- Las variables deben haberse medido por lo menos en una escala de intervalo de manera que sea posible utilizar las operaciones aritméticas.

Referencias

[1] Flores-Ruiz, E., Miranda-Novales, María, Villasís-Keever, Miguel. The research protocol VI: How to choose the appropriate statistical test. Inferential statistics. *Revista Alergia Mexico*, 64:364–370, 07 2017.

[2] Hernandez, Freddy. Prueba de hipótesis.

[3] INEGI. Índice Nacional de Precios al Consumidor (INPC). Recurso libre, disponible en https://www.inegi.org.mx/datos/, 2020.

[4] Prabhakaran, Selva. Statistical Tests. Recurso disponible en:http://r-statistics.co/Statistical-Tests-in-R.html, 2017.

[5] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.

[6] Walpole, Ronald E., Myers, Raymond H., Myers Sharon L., Ye, Keying. *Probability statistics for Engineers Scientists*. Pearson Education, Inc., 9th edition, 2012.